Volume 9, Special Issue No. 1, February 2018

ISBN: 978-93-5300-455-2

International Journal of Advanced Research in Computer Science
(ISSN: 0976-5697)
RESEARCH PAPER

ICRACC 2018
(23rd–24th February 2018)

Available Online at www.ijarcs.info

Semantic Similarity Measures: an Overview and Comparison

Dr.B.Poorna,
Principal, Shri.Shankarlal Sundarbai Shasun Jain college For Women, Chennai, India
poornasundar@yahoo.com

A. Sudha Ramkumar,
Research Scholar, Bharathiar University
Coimbatore, India
sudharam99@gmail.com

ABSTRACT

Similarity measure is calculated based on the syntactical representation of terms. Similarity measure used in data mining task likes clustering, and classification returns irrelevant information. The semantic similarity calculated based on the relatedness between wordpairs of terms returns better result. Many researchers proposed approaches for getting word similarity by using different sources like ontologies, thesauri etc. This Paper provides an overview of six existing semantic similarity measures and compared those semantic similarity measures using the wordpairs of sports domain ontology. This paper describes about how WordNet is used to retrieve the synonyms and using synsets how semantic similarity measures are calculated. Finally, comparison of selected semantic similarity measures for the given wordpairs with respect to the considered knowledge base domain ontology and WordNet is presented.

***Keywords*:** Similarity measure; WordNet; Synonyms;

I. Introduction

In the digital era, the amount of electronic document has been increasing tremendously. The documents retrieved as a result of search query depends only on the terms. Hence, the retrieved result consists of a combination of both relevant and irrelevant content. If the machine understands the user requirement, it will fetch relevant content. In order to make machine understandable, ontologies become an integral part in today's information retrieval. Semantic Similarity measure is calculated based on the likeliness of the term's meaning. Many researchers proposed knowledge sources like WordNet Ontology in their work to prove how it can be used to calculate the semantic similarity between terms or concepts of ontology. WordNet is the lexical database developed at Princeton University and can be interpreted and used as ontology in the computer science. It is an online database which includes nouns, verbs, adjectives and adverbs grouped into sets of synonyms called synsets. Many researchers proposed that WordNet is widely used to compute the semantic similarity measure between the concepts and it reduces the dimensionality of the term matrix.

In this paper, an overview of six existing semantic similarity measures like wu&palmer, Leacock & Chodorow, HirstOnstonge, Resnik, Jiang & Conrath and Lin measures is provided. An experimental result shows the comparison of these six measures for the word pairs of sports domain ontology along with the WordNet. As a first step of this study an Overview of six existing semantic similarity measure is presented, followed by comparison of six similarity measure for the wordpairs of sports domain. In this study, new synonym retrieval algorithm is implemented to retrieve synonyms of all the selected terms using WordNet ontology and finally the similarity measure calculated to select only the most relevant synonym of the terms. This paper proves experimentally the performance of the six existing similarity measure. This paper is organized as follows Section 2 provides the Literature review of semantic similarity measure. Section 3 presents the overview of six semantic similarity measures. Section 4 compares the six semantic similarity measures followed by Conclusion in Section 5.

II. Literature Review

Zhang et al, presented a comparative study on different semantic similarity measures of term including path based measure, information content based measure and feature based semantic similarity measure affect document clustering. In their article, the domain ontology is integrated with the clustering process by reweighting the terms and proved that it has positive effects on document clustering. Mesh Ontology is used as knowledge source in this paper [15]. Zhang et al, presented nine semantic similarity measures with a term reweighting method on PubMed document. The experimental result shows that term reweighting has some positive effects on clustering and proved path based semantic similarity measures improves the performance significantly. Domain ontology is acting as a knowledge source in this paper [16].

Montserrat Batet et al. analyzed the existing semantic similarity measures by determining their advantage

and limitation based on the knowledge base. This paper proposed a new measure based on the exploitation of the taxonomical structure. SNOMED CT Ontology is used as knowledge source and accuracy of their proposed measure is compared with existing measure [2]. Lingling Meng et al. presented an effective algorithm for semantic similarity metric of word pairs. This new algorithm considers both path length and information content. This proposed algorithm outperformed traditional similarity algorithm. Here, the WordNet ontology is used as a knowledge source [10].

Gan et al, classified existing semantic similarity calculation method into 5 categories as Based on semantic distance, based on Information content, based on properties of terms, based on ontology hierarchy and hybrid method. In this paper, the knowledge resource is domain ontology. Finally, they provided a summary of characteristics, advantage and disadvantage of each category. Finally, concluded that these methods depend on 2 factors – the quality of annotation data and the correct interpretation of the hierarchical structure of ontology [4]. Thabet slimani et al, discussed about the existing semantic similarity measures based on path, information content and feature based. Based on two standard benchmarks, a calculation of all approaches is presented [12].

Mabotuwana et al, presented a semantic vector based approach to determine similarity between documents using domain ontology. This semantic algorithm improves classification accuracy when compared to non-semantic approach. Here, the domain ontology is used as a source [9]. Cui et al, proposed WordNet based semantic similarity clustering algorithm on the cluster analysis of complex network community. The proposed algorithm is compared with VSM and K-Means and proved with effective result. Here, the WordNet ontology is the knowledge resource [3]. Ali Hadj et al, proposed a new measure which combines the most significant parameters depth and hyponym of a concept. Experimental result shows that proposed measure outperforms existing path based, information content based and feature based approaches. The knowledge source used here is WordNet [13]. Ahmad Fayez et al, focuses only on the semantic similarity measure based on ontology as a knowledge source. In this paper, Al-Mubaid &Nyguan's method outperforms the other measure [1].

III. An Overview

In this section, an overview of six existing semantic similarity measures is provided. The semantic similarity measures considered in this section belongs to path based measure and information content based measure. The path based measures discussed here are wu&palmer measure, Leacock&Chodorow, Hirst &On-Sage measure and path based semantic similarity measure. The information content based measures considered for the comparison in this paper are Resnik semantic measure and Lin Semantic measure.

3.1 Path Based Measure

Wu et al, 1994 proposed a path based measure [14] that considers the depth of the concepts in the hierarchy. This measure calculates the similarity value by considering the depth of the two synsets in the WordNet, along with the depth of the least common subsumer. The wu&palmer measure ranges from 0 to 1.

$$SS_{W\&P} = \frac{2*\text{Depth (LCS)}}{\text{Depth(S1)+Depth(S2)}} \quad (1)$$

Leacock and chodorow [7] proposed a path based measure that depends on the length (C1, C2) of the shortest path between two synsets or wordpairs for their similarity measure. This measure considers IS-A links and scale the path length by the overall depth D of the taxonomy. The leacock and chodorow values lies from 0 to 4.

$$SS_{L\&C} = -\log \left(\frac{SP\ (C1, C2)}{2(max\_depth)} \right) \quad (2)$$

Hirst-St-Onge Measure [5] is a measure of semantic relatedness in that two lexicalized concepts are semantically close if their WordNet synsets are connected by a path are connected by a path that is not too long and that "does not change direction too often".

$$SS_{HS} = C - \text{Path Length} - K \times D \quad (3)$$

D is the number of changes of direction in the path. C and K are the constants. If SS_{HS} is zero then there is no path exists in between the concepts. The Hirst-St-Onge values ranges from 0 to 16.

3.2 Information Content Based Measure

Resnik similarity measure [11] considers the integration of ontology and corpus. Resnik defined the similarity between two concepts lexicalized in WordNet to be the information content of their lowest super coordinate that is the most specific common subsumer.

$$SS_{Res} = -\log P\ (LCS\ (C1, C2)) = IC\ ((LCS\ (C1, C2)) \quad (4)$$

Where $IC(C) = \frac{\log (\text{depth}(C)}{\text{Log}(\log(\text{deep}_{max})}$

Jiang and Conrath measure [6] is based on information content. Here, the distance between two concepts c1 and c2 is calculated as the difference between the sum of information content of the two concepts c1 and c2 and the information content of their most informative

subsume. Jiang & conrath measure is calculated using the following formula,

$$SS_{JC} = 2 * \ln P_{mis}(C1, C2) - (\ln(P(C1) + \ln(P(C2)) \quad (5)$$

This measure is the shortest path length between two concepts c1 and c2 and the density of concepts along the same path.

Lin Similarity measure [8] follows from his theory of similarity between arbitrary objects. It uses the same element as Jiang and Conrath. It is based on Resnik's similarity and it considers both the information content of lowest common subsume and two compared concepts. The Lin Similarity measure is calculated as follows,

$$SS_{Lin} = \frac{2 * sim_{Res}(C1, C2)}{IC(C1) + IC(C2)} \quad (6)$$

IV. COMPARISON

We conduct experiments on bbcsport dataset. We developed sports domain ontology using protégé tool. The concepts of sports domain ontology are extracted using Jena, a java framework for OWL ontology. The extracted concepts are mapped with the extracted terms of bbcsports dataset. After the term-concept match, the terms are selected for further processing. The selected terms are searched in WordNet for the synonyms. The synonyms returned consist of an array of words and is called as synsets. The pair of words of synsets is applied to the six semantic similarity measures. The similarity measures are calculated with the help of WordNet Similarity for Java (WS4J).

The experiment is conducted in eclipse luna, a integrated development environment. In order to compare six semantic similarity measures, the values are normalized to value ranges from 0 to 1. For the wordpair "centuries" and "hundred", the following table shows the semantic similarity values,

Table 1: Comparison of Six Semantic Similarity Value

Category	*Measure*	*Semantic similarity Value*	*Normalized value*
Path Based	*Wu&Palmer*	*1.0*	*1.0*
	Leacock&Chodorow	*3.688*	*0.92*
	Hirst and St-Onge	*16*	*1.0*
Information content based	*Resnik*	*8.4699*	*0.94*
	Jiang and Conrath	*1.2876*	*0.85*
	Lin	*1.0*	*1.0*

The value relies on the knowledge source like WordNet, thesauri and Domain ontology. If the knowledge source is domain ontology, then the value of the semantic similarity depends upon the correct interpretation of concept hierarchy. If the taxonomy is not correct, then there will be misinterpreted value will be returned as a result. So care must be taken during the development of domain ontology. If the WordNet is the knowledge source, then the value will be accurate. The following fig.1 shows the comparison of six semantic similarity values.

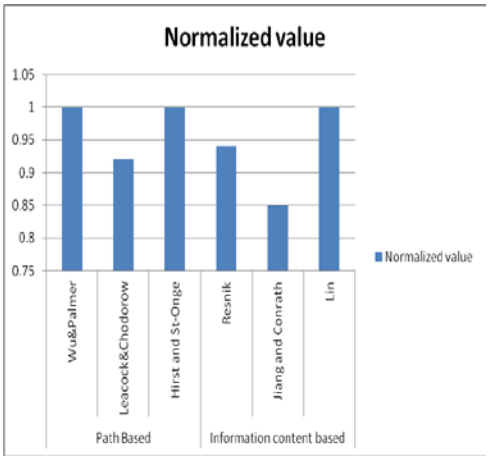

Figure.1 Comparison of six similarity measures

V. CONCLUSION

This paper compared six selected semantic similarity measure. The experimental result shows that Wu&Palmer, Hirst-St-Onge measure and Lin measure outperforms other semantic similarity measure. The Wu & Palmer and Hirst-St-Onge measure belongs to the path based category whereas the Lin measure belongs to the information content based measure. These measures play an important role in document clustering and classification process, because they reduce the complexity of clustering process by reducing the dimensionality of the term matrix. There are many other semantic similarity measures exists, depends upon the application and the knowledge source, the similarity measure improves the performance of the clustering process. This paper attempts to help the researcher to understand about the importance of semantic similarity measure in clustering process.

VI. REFERENCES

[1] Althobaiti, A. F. S. (2017). Comparison of Ontology-Based Semantic-Similarity Measures in the Biomedical Text. Journal of Computer and Communications, 5(02), 17.

[2] Batet, M., Sánchez, D., & Valls, A. (2011). An ontology-based measure to compute semantic similarity in biomedicine. Journal of biomedical informatics, 44(1), 118-125.

[3] Cui, L. Z., Lu, N., & Jin, Y. Y. (2014). Community Clustering Algorithm on Semantic Similarity in Complex Network. Lecture Notes on Software Engineering, 2(4), 348.

[4] Gan, M., Dou, X., & Jiang, R. (2013). From ontology to semantic similarity: calculation of ontology-based semantic similarity. The Scientific World Journal, 2013.

[5] Hirst, G., & St-Onge, D. (1998). Lexical chains as representations of context for the detection and

correction of malapropisms. WordNet: An electronic lexical database, 305, 305-332.

[6] Jiang, J. J., & Conrath, D. W. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. arXiv preprint cmp-lg/9709008.

[7] Leacock, C., & Chodorow, M. (1998). Combining local context and WordNet similarity for word sense identification. WordNet: An electronic lexical database, 49(2), 265-283.

[8] Lin, D. (1998, July). An information-theoretic definition of similarity. In Icml (Vol. 98, No. 1998, pp. 296-304).

[9] Mabotuwana, T., Lee, M. C., & Cohen-Solal, E. V. (2013). An ontology-based similarity measure for biomedical data–Application to radiology reports. Journal of biomedical informatics, 46(5), 857-868.

[10] Meng, L., Huang, R., & Gu, J. (2013). An effective algorithm for semantic similarity metric of word pairs. International Journal of Multimedia and Ubiquitous Engineering, 8(2), 1-12.

[11] Resnik, P. (1999). Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. J. Artif. Intell. Res.(JAIR), 11, 95-130.

[12] Slimani, T. (2013). Description and evaluation of semantic similarity measures approaches. arXiv preprint arXiv:1310.8059.

[13] Taieb, M. A. H., Aouicha, M. B., & Hamadou, A. B. (2014). Ontology-based approach for measuring semantic similarity. Engineering Applications of Artificial Intelligence, 36, 238-261.

[14] Wu, Z., & Palmer, M. (1994, June). Verbs semantics and lexical selection. In Proceedings of the 32nd annual meeting on Association for Computational Linguistics (pp. 133-138). Association for Computational Linguistics.

[15] Zhang, X., Jing, L., Hu, X., Ng, M., & Zhou, X. (2007, April). A comparative study of ontology based term similarity measures on PubMed document clustering. In International Conference on Database Systems for Advanced Applications (pp. 115-126). Springer, Berlin, Heidelberg.

[16] Zhang, X., Jing, L., Hu, X., Ng, M., Xia, J., & Zhou, X. (2008). Medical document clustering using ontology-based term similarity measures.